Current **A**pplied **S**cience and **T**echnology e-ISSN: 2586-9396

Curr. Appl. Sci. Technol. 2027, Vol. 27 (No. 1), e0267953

Research article

Nanotoxicity Prediction Based on Half Maximal Effective Concentration (EC50) Values Using Elastic Net Regression and k-Nearest Neighbors

Erica Denise V. Arceno[1], Aries P. Valeriano[1]*, John Riz V. Bagnol[1], Farlley G. Bondaug[2,3], Irelie P. Ebardo[2,3], Penelope P. Almonte[3], Michelle Amor P. Sabugaa[3] and Rey Y. Capangpangan[4]

[1]Department of Mathematics and Statistics, University of Southeastern Philippines, Davao City, Philippines
[2]Department of Science and Technology - Science Education Institute, Taguig City, Philippines
[3]Research on Environment and Nanotechnology Laboratories, Research Division, Mindanao State University at Naawan, Naawan, Philippines
[4]Department of Physical Sciences and Mathematics, Mindanao State University at Naawan, Naawan, Philippines

Received: 8 June 2025, Revised: 3 December 2025, Accepted: 23 December 2025, Published: 24 March 2026

Abstract

This study focuses on predicting the toxicity of nanoparticles using their half maximal effective concentration (EC50) values through Elastic Net Regression (ENET) and k-Nearest Neighbors (kNN) models. A dataset of 5,926 samples with 10 descriptors was used to develop Nano - Quantitative structure - activity relationship (Nano-QSAR) models. The kNN model with $k = 2$ demonstrated the best performance, achieving an R^2 of 91.66%, RMSE of 0.79, and MAE of 0.34 on the training set, and an R^2 of 91.58%, RMSE of 0.79, and MAE of 0.32 on the test set, indicating strong predictive ability and good generalization to new data. The ENET model, however, showed substantially lower accuracy with performance measures below 60%, suggesting it was not suitable for this dataset. Key descriptors associated with electronegativity and atomic properties were identified as major contributors to toxicity. These findings highlight the potential of kNN as an effective approach for predicting nanoparticle toxicity and supporting the development of safer nanomaterials.

Keywords: Elastic Net Regression; k-Nearest neighbors; nanoparticles; toxicity; Nano-QSAR models

*Corresponding author: E-mail: apvaleriano@usep.edu.ph
https://doi.org/10.55003/cast.2026.267953
Copyright © 2026 by King Mongkut's Institute of Technology Ladkrabang, Thailand. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Nanotechnology involves redesigning and manipulating matter at the atomic level to develop new materials and devices (Roco et al., 1999). As defined by the National Nanotechnology Initiative (NNI), it deals with particles that are one-billionth of a meter in size, which help improve and revolutionize technology and medicine. According to several studies (Baig et al., 2021; Armarego, 2022), nanomaterials are applied in various aspects of society and fields of study such as health, water, and consumer products, and have even been used unknowingly since ancient times. The growing application of nanomaterials in product development has driven the advancement of toxicity assays and physicochemical characterization methods to assess their potential toxicity (Ahadian & Radisic, 2017). Understanding nanomaterials through multiple approaches is crucial as it may help develop mitigation strategies that encourage the safe application of nanoparticles (Wen et al., 2021).

Nanotechnology has been applied in many disciplines. One example is civil engineering, which uses nanomaterials in design and construction processes. Through nanotechnology, lighter structures, stronger structural composites for bridges and other infrastructure, and lower maintenance coatings have become possible (Gajanan & Tijare, 2018). In medicine, nanotechnology has been applied in the development of treatments for diseases like cancer by targeting therapeutic agents directly to tumor cells, effectively destroying cancerous cells while minimizing harm to healthy tissue (Misra et al., 2010; Gajanan & Tijare, 2018). Nanoparticles can encapsulate drugs, genes, and imaging agents, enhancing their solubility, stability, and ability to cross biological barriers (Kou et al., 2018).

The impressive and effective use of nanoparticles in nanotechnology has led to many advancements and promises more as the field matures. However, comprehending and addressing the potential risks of nanotechnology to human health and the environment continues to pose significant challenges (Beaudrie, 2011). While the long-term health effects of prolonged exposure to varying concentrations of nanoparticles are not yet fully understood, the environmental impact of nanomaterials is anticipated to grow in the coming years.

Nanotoxicology aims to understand the molecular mechanisms underlying the bioaccumulation and toxicity of nanoparticles (AshaRani et al., 2009). It plays a crucial role in identifying the potential harmful effects of nanoparticles and nanopharmaceuticals on both human health and the environment (Erkekoglu & Kocer-Gumusel, 2018). Quantitative structure-activity relationship (QSAR), formulated by Hansch et al. (1962), is one of the most promising methods that could supply missing data regarding nanotoxicity, which is crucial for risk assessment. It employs computational chemistry based on the premise that variations in molecular structure influence a compound's biological activity (Puzyn et al., 2009). For instance, a QSAR model was applied to predict the cytotoxicity of metal oxide nanoparticles, yielding results aligned with findings from several studies cited in the paper: zinc oxide, copper oxide, nickel oxide, and cobalt oxide nanoparticles exhibited the highest levels of cytotoxicity, whereas titanium oxide was found to be the least toxic (Puzyn et al., 2011).

Given the complexity of nanotoxicity, there is an urgent need to develop methods to predict toxicity and identify the properties and behavior of nanoparticles (Li et al., 2022). This study aims to develop and assess the performance of Elastic Net Regression (ENET) and k-Nearest Neighbors (kNN) algorithms in predicting nanoparticle toxicity based on median effective concentration (EC50), following the guidelines set by the Organization for

Economic Co-operation and Development (OECD, 2014). The research contributes to advancing the understanding of nanotoxicity and supports the safe and responsible use of nanoparticles across diverse applications.

2. Materials and Methods

2.1 Source of data

This study utilizes data from Concu et al. (2017), comprising 5,926 observations with ten molecular descriptors and one response variable (EC50). In their work, they applied a perturbation theory approach in which each nanoparticle (NP) is evaluated relative to another through case–case pairs, where one NP serves as the reference state (baseline with known properties) and the other as the new state (the NP being predicted). The descriptors quantify the perturbations, or differences, between these paired states across structural, physicochemical, and toxicological dimensions. The ten descriptors capture variations in these properties between nanoparticles in the new and reference states. $TE_i(c_j)_{rf}$ is a dummy variable encoding the toxic effect of the reference NP. $DDV(m_e)$ quantifies changes in molar volume, while $DDL(m_e)$ reflects variations in particle size, both dependent on toxicity measurements. Spectral moment-based descriptors include $DD\mu_1(ATO)b_t$, which measures differences in the first-order spectral moment weighted by atomic weight and influenced by biological systems, and $DD\mu_3(POL)n_s$, which accounts for changes in the third-order spectral moment weighted by polarizability, and associated with nanoparticle shape. $DDE(d_m)$ represents shifts in electronegativity between nanoparticles, conditioned by the measurement environment of particle size. Time-dependent descriptors include $DD\mu_3(VAN)t_a$, which captures differences in the third-order spectral moment weighted by the van der Waals radius, and $DD\mu_2(ATO)_{ta}$, which reflects changes in the second-order spectral moment weighted by atomic weight, both dependent on exposure duration. Finally, coating-related descriptors include $DG\mu_2(Hyd)s_c$, a general second-order spectral moment weighted by hydrophobicity, and $DG\mu_5(PSA)s_c$, a fifth-order general spectral moment weighted by polar surface area, both describing structural differences in coating agents between the new and reference NPs.

2.2 Software used

This study was conducted using R, an open-source programming language for statistical computing, visualization, and model development. Data preprocessing, including standardization and transformation, was performed using the recipes package (Kuhn et al., 2024). The rsample package was used to split the dataset into training and testing sets (Frick et al., 2023). Model training, hyperparameter tuning, model evaluation, and variable importance calculation were carried out with the caret package (Kuhn, 2008). The Tidyverse packages were used for data organization (Wickham et al., 2019). For visualization, ggplot2 (Wickham, 2016) was used to create plots of model performance, and patchwork (Pedersen, 2025) was used to arrange multiple plots in a single figure.

2.3 Machine learning algorithms

2.3.1 Elastic Net Regression (ENET)

Elastic Net is a regularized linear regression method that integrates the advantages of both LASSO and ridge regression. It is especially useful for datasets with numerous predictors or high multicollinearity. By applying a blend of L1 (LASSO) and L2 (ridge) penalties to the model coefficients, Elastic Net enables variable selection and simultaneously shrinks coefficients to reduce the risk of overfitting. The amount of regularization is controlled by λ, a hyperparameter that must be tuned (Zou & Hastie, 2005; Boehmke & Greenwell, 2019).

2.3.2 k-Nearest Neighbors (kNN)

The k-Nearest Neighbors (kNN) algorithm is a straightforward machine learning method that estimates the value of a new observation by averaging the outcomes of the k most similar instances within the dataset. Similarity is usually measured using distance metrics such as Euclidean distance. The method does not build an explicit model but relies on the data at the time of prediction, which can be computationally demanding. The value of k impacts model performance: small k may cause overfitting, while large k can lead to underfitting (Cunningham & Delany, 2007; Boehmke & Greenwell, 2019).

2.4 Procedure of the study

2.4.1 Data splitting

The data was split using an 80:20 ratio, where 80% was used for model development and 20% for external validation, following Gholamy et al. (2018).

2.4.2 Data preprocessing

Data preprocessing involves cleaning and organizing the data to make it suitable for model development. In this phase, the descriptors underwent standardization to improve data consistency, as the predictor variables have different units of measurement. This step helps produce more reliable results. The response variable was also transformed using the Yeo-Johnson transformation to address possible outliers and to approximate normality (Zou & Hastie, 2005).

2.4.3 Model building

Models were built on the training set and evaluated for robustness using 10-fold repeated cross-validation. Grid search was applied to optimize hyperparameters. For k-Nearest Neighbor (kNN), k values from 2 to 25 were tested. For Elastic Net Regression (ENET), the regularization parameter λ ranged from 0.00001 to 0.25. The mixing parameter α was set to 0.5, which gave equal weight to LASSO (L1) and Ridge (L2) penalties (Boehmke & Greenwell, 2019). This allowed the model to both select the most important descriptors (LASSO) and shrink the coefficients to avoid overfitting (Ridge). This choice improved predictive accuracy while maintaining model stability and efficiency. Model performance was assessed using R^2, RMSE, and MAE metrics.

2.4.4 External validation

After tuning the models to find the optimal hyperparameters, their performance was externally validated on the testing set. This step assesses the models' generalizability and performance on new, unseen data and is essential for evaluating their practical utility in real-world applications.

2.4.5 Variable importance

The most influential descriptors on the response variable were identified using different approaches for each model. For elastic net regression, variable importance was determined based on the absolute values of the model's coefficients. Variables with larger absolute coefficients are considered more influential in predicting the outcome. For k-nearest neighbors (kNN), variable importance was assessed using recursive feature elimination (RFE). This method evaluates model performance across different subsets of variables, starting from the smallest set and gradually increasing in size. The variables that were selected earlier in the process and had consistently improved model performance (R^2) were considered the most important.

2.4.6 Applicability domain

Applicability domain analysis was conducted using the PCA bounding box method to ensure compliance with OECD guidelines. This analysis helps determine whether new samples fell within the model's reliable prediction space (Phanus-umporn et al., 2018).

3. Results and Discussion

3.1 The Nano-QSAR models internal validation and hyperparameter tuning

To identify the optimal hyperparameters, 10-fold repeated cross-validation was performed for both Elastic Net Regression (ENET) and k-Nearest Neighbors (kNN). The results of this tuning are shown in Figures 1 and 2 respectively, where RMSE, MAE, and R^2 values were shown across different λ and k values. To better understand the performance of each model, the results are discussed separately, beginning with ENET followed by kNN.

Figure 1 shows the training performance of ENET model across different values of λ. The results indicate that the optimal λ was 0.00311, achieving an RMSE of 1.80 and an R^2 of 56.67%. Although the MAE at this λ (1.42) was not the lowest, the chosen λ minimized RMSE and maximized R^2, which justifies its selection. This means that ENET could explain just over half of the variance in nanoparticle toxicity based on EC50, but its relatively high RMSE and MAE suggested that its predictions were often far from the true EC50 values, limiting its effectiveness.

Figure 2 illustrates the training performance of kNN across varying k values. The optimal value was k = 2, yielding an RMSE of 0.79, an MAE of 0.34, and an R^2 of 91.66%. These metrics show that the kNN predictions were much closer to the actual EC50 values compared to ENET, while also explaining over 91% of the variability in nanoparticle toxicity. For k > 2, both RMSE and MAE increased while R^2 declined, suggesting reduced accuracy as less similar neighbors were included. Thus, k = 2 provided the best trade-off, effectively

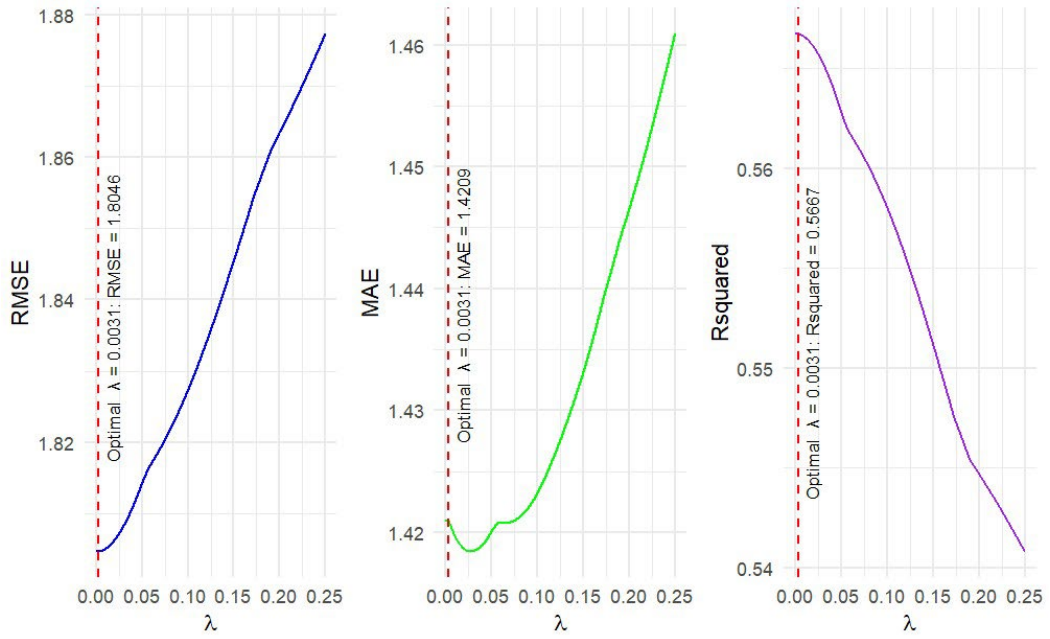


Figure 1. Training performance of the ENET Model across different λ values

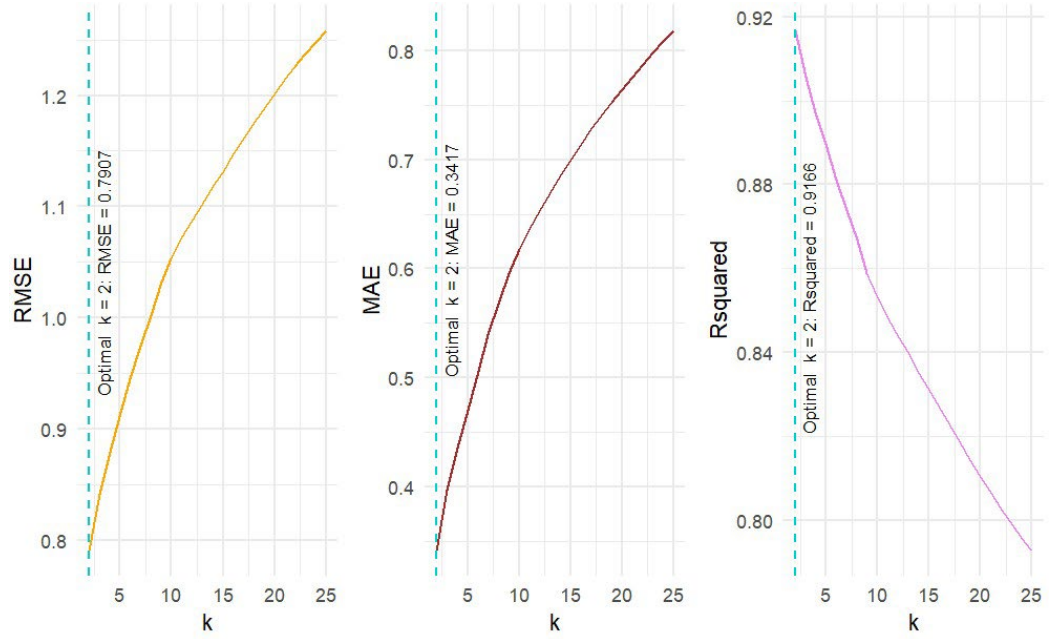


Figure 2. Training performance of the kNN Model across different k values

capturing descriptor–toxicity relationships (e.g., atomic and electronegativity features) without being overly influenced by noise. These results confirmed that the chosen hyperparameters were indeed optimal and that kNN substantially outperformed ENET in predictive accuracy and consistency. While using a small k (such as 2) is typically associated with a risk of overfitting, this concern is directly addressed in Section 3.2, where external validation demonstrates that the kNN model maintained its high performance on unseen data, confirming its robustness and reliability.

3.2 The optimal Nano-QSAR models external validation

To clearly show the predictive ability of the developed models, Table 1 presents the first ten true EC50 values from the test set alongside their corresponding predictions from ENET and kNN. The results reveal that kNN predictions were consistently close to the true values, with only small deviations (e.g., true value 5.13 vs. kNN prediction 5.06). In contrast, ENET often showed larger discrepancies (e.g., true value 0.64 vs. ENET prediction 4.46), indicating weaker predictive accuracy. These examples provide concrete evidence that kNN was more reliable in capturing the underlying patterns in the data compared to ENET.

Table 1. First ten ec50 true values with corresponding predictions from Enet and kNN

True Value	ENET Predicted Value	kNN Predicted Value
5.13	4.76	5.06
5.68	6.83	5.21
0.64	4.46	0.64
8.83	7.27	8.83
2.11	2.33	2.11
0.38	1.06	0.38
5.13	3.52	5.08
4.15	5.14	3.97
2.11	2.64	2.11
5.13	4.08	5.06

Note: The supporting document with the file name “predicted values” provides the complete set of predicted values from ENET and kNN.

The sample comparisons in Table 1 are consistent with the overall summary statistics shown in Table 2, where kNN clearly outperformed ENET across all external validation metrics. Specifically, kNN achieved an RMSE of 0.79, an R^2 of 91.58%, and an MAE of 0.32, confirming its high predictive accuracy and generalizability. In contrast, the ENET model showed lower performance with an RMSE of 1.77, an R^2 of 58.17%, and an MAE of 1.40. Comparing these results to the internal validation in Section 3.1, both models exhibit similar performance on unseen data, with minimal changes in error metrics and explained variance. This consistency indicated that overfitting was avoided. The kNN model, in particular, maintained its strong predictive ability, confirming its robustness and reliability for predicting nanoparticle toxicity based on EC50 values.

Table 2. External validation performance of the optimal Nano-QSAR models

Model	RMSE	R^2(%)	MAE
ENET	1.77	58.17	1.40
kNN	0.79	91.58	0.32

3.3 The analysis of variable importance

Table 3 presents the variable importance results for ENET and kNN. For ENET, descriptors were ranked according to the absolute values of their regression coefficients ($|\beta|$), where a higher $|\beta|$ indicated greater influence. For kNN, importance was measured by the increase in R^2 when descriptors were added sequentially. The R^2 values in Table 3 therefore applied only to kNN and reflected the stepwise contribution of each descriptor to model performance. For ENET, the top three descriptors were $DD\mu_1(ATO)b_t$ ($|\beta| = 1.43$), which characterized the change in the spectral moment of order 1 weighted by atomic weight, reflecting how NP structure interacted with different biological systems; $DDE(d_m)$ ($|\beta| = 1.02$), representing variations in electronegativity between NPs depending on how their sizes were measured, emphasizing the importance of electronic properties in toxicity prediction; and $DG\mu_2(Hyd)s_c$ ($|\beta| = 0.69$), which quantified differences in hydrophobicity between the chemical structures of NP coatings, underscoring the role of surface chemistry in NP behavior. Among these, $DD\mu_1(ATO)b_t$ and $DDL(m_e)$ stood out with substantially higher weights, while $DG\mu_2(Hyd)s_c$ had a comparatively smaller effect. In contrast, kNN showed a broader pattern of descriptor contributions. Its top three descriptors were $DDE(d_m)$, $DDL(m_e)$, and $DD\mu_1(ATO)b_t$, which together raised the explained variance from 64.69% to 80.60%. The inclusion of $DDE(d_m)$ among the top three reflected the importance of size-related perturbations. Beyond these, additional descriptors provided smaller but consistent improvements, with the full set achieving an R^2 of 88.74%. A direct comparison shows both similarities and differences between the models. ENET and kNN both identified $DDE(d_m)$ and $DD\mu_1(ATO)b_t$ among their top three descriptors, while kNN also emphasized $DDL(m_e)$, which ENET ranks fourth. It is important to note that the R^2 values in Table 3 were not directly comparable to those reported in Figures 1 and 2 or Table 2. Figures 1 and 2 show the overall R^2 of the hyperparameter tuned ENET and kNN models under 10-fold repeated cross-validation (ENET with 56.67% and kNN with 91.66%), while Table 2 presents their external validation on the test set. By contrast, Table 3 shows only the incremental R^2 values from kNN's variable importance analysis, with the maximum of 88.74% slightly below the final values in Figure 2 and Table 2. This was expected, as the optimized kNN model achieved its best performance through the combined effect of all descriptors. Together, these findings reveal a key difference in how the models rely on descriptors. ENET depended mainly on two descriptors with strong influence ($DD\mu_1(ATO)b_t$ and $DDE(d_m)$), while the others contribute little, limiting its performance. kNN, in contrast, gained substantial improvements from its top three descriptors and further enhanced accuracy through smaller contributions from additional ones. This broader accumulation of information explained the higher predictive accuracy of kNN in Figures 1 and 2 and Table 2.

Table 3. Variable importance results for ENET and kNN models

ENET		kNN	
Descriptor	$\|\boldsymbol{\beta}\|$	**Descriptor/s**	$\boldsymbol{R^2}$ **(%)**
$DD\mu_1(ATO)b_t$	1.43	$DDE(d_m)$	64.69
$DDE(d_m)$	1.02	$DDE(d_m) + DDL(m_e)$	73.10
$DG\mu_2(Hyd)s_c$	0.69	$DDE(d_m) + DDL(m_e) + DD\mu_1(ATO)b_t$	80.60
$DDL(m_e)$	0.54	$DDE(d_m) + DDL(m_e) + DD\mu_1(ATO)b_t + DG\mu_5(PSA)s_c$	82.14
$DD\mu_3(VAN)t_a$	0.54	$DDE(d_m) + DDL(m_e) + DD\mu_1(ATO)b_t + DG\mu_5(PSA)s_c + DDV(m_e)$	84.64
$DD\mu_3(POL)n_s$	0.47	$DDE(d_m) + DDL(m_e) + DD\mu_1(ATO)b_t + DG\mu_5(PSA)s_c + DDV(m_e) + DG\mu_2(Hyd)s_c$	84.52
$DD\mu_2(ATO)t_a$	0.35	$DDE(d_m) + DDL(m_e) + DD\mu_1(ATO)b_t + DG\mu_5(PSA)s_c + DDV(m_e) + DG\mu_2(Hyd)s_c + DD\mu_3(VAN)t_a$	86.13
$DG\mu_5(PSA)s_c$	0.32	$DDE(d_m) + DDL(m_e) + DD\mu_1(ATO)b_t + DG\mu_5(PSA)s_c + DDV(m_e) + DG\mu_2(Hyd)s_c + DD\mu_3(VAN)t_a + DD\mu_2(ATO)t_a$	87.26
$TE_i(c_j)_{_rf}$	0.27	$DDE(d_m) + DDL(m_e) + DD\mu_1(ATO)b_t + DG\mu_5(PSA)s_c + DDV(m_e) + DG\mu_2(Hyd)s_c + DD\mu_3(VAN)t_a + DD\mu_2(ATO)t_a + DD\mu_3(POL)n_s$	87.41
$DDV(m_e)$	0.07	$DDE(d_m) + DDL(m_e) + DD\mu_1(ATO)b_t + DG\mu_5(PSA)s_c + DDV(m_e) + DG\mu_2(Hyd)s_c + DD\mu_3(VAN)t_a + DD\mu_2(ATO)t_a + DD\mu_3(POL)n_s + TE_i(c_j)_{_rf}$	88.74

3.4 Applicability domain assessment

Figure 3 shows the results of applying the PCA bounding box, where the training set is represented by blue points and the testing set by red points. It can be observed that most of the testing set data points are positioned close to the training set data points, with no clear separation between the two sets. This suggests that the testing set falls within the boundary of the training set. Therefore, the model built using the training set is applicable to the testing set, which validates the model evaluation results presented earlier.

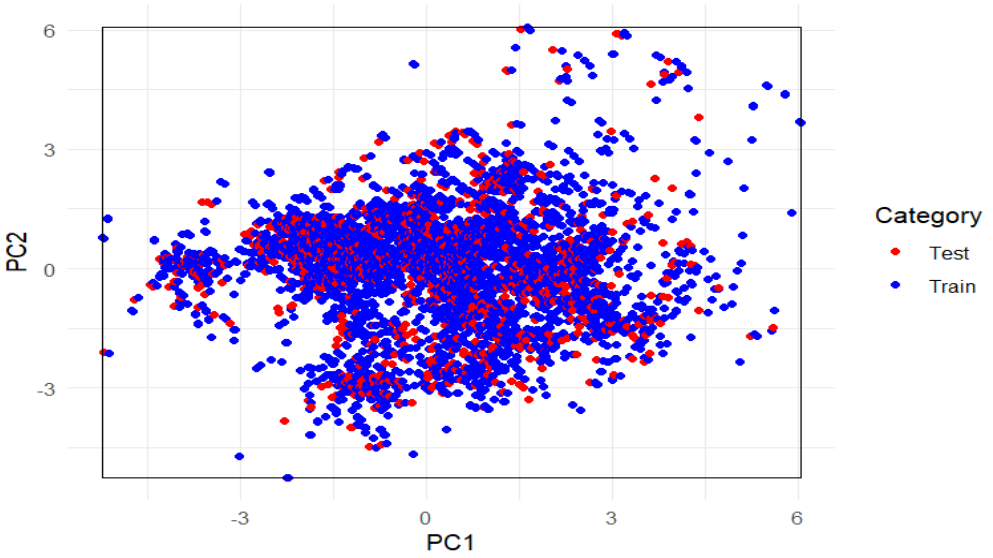


Figure 3. PCA bounding box

4. Conclusions

This study successfully developed and evaluated Nano-QSAR models to predict nanoparticle toxicity based on EC50 values using Elastic Net Regression (ENET) and k-Nearest Neighbors (kNN) algorithms. Among the two approaches, the kNN model (with $k = 2$) exhibited a stronger predictive performance, achieving an R^2 of approximately 91.66% and low error values (RMSE = 0.79, MAE = 0.34), indicating its suitability for modeling nonlinear patterns in nanotoxicity data. Analysis of variable importance revealed that descriptors associated with electronegativity ($DDE(d_m)$) and atomic properties ($DD\mu_1(ATO)b_t$) significantly contributed to toxicity prediction, emphasizing the relevance of electronic and atomic-level characteristics. Additionally, PCA bounding box applicability domain analysis confirmed that the test data fall within the model's reliable prediction space. Overall, the results indicate that the kNN algorithm provides a reliable and effective computational method for assessing nanoparticle safety, highlighting its potential to support the responsible development and use of nanomaterials in various fields.

5. Acknowledgements

The authors acknowledge the support of the Department of Science and Technology - Philippine Council for Industry, Energy, and Emerging Technology Research and Development (DOST-PCIEERD) under Grant No. 10185.

6. Authors' Contributions

Erica Denise V. Arceno and Aries P. Valeriano: designed research, performed research, wrote the paper. John Riz V. Bagnol and Farlley G. Bondaug: contributed analytic tools. Penelope P. Almonte and Michelle Amor P. Sabugaa: reviewing. Irelie P. Ebardo: editing; and Rey Y. Capangpangan: funding acquisition, project administration.

ORCID
Aries P. Valeriano https://orcid.org/0000-0003-4029-0666
John Riz V. Bagnol https://orcid.org/0009-0004-9952-4310
Penelope P. Almonte https://orcid.org/0000-0002-9778-7086
Michelle Amor P. Sabugaa https://orcid.org/0009-0001-2452-4306
Rey Y. Capangpangan https://orcid.org/0000-0002-7078-9903

7. Conflicts of Interest

The authors declare that there are no conflicts of interest related to this work.

References

Ahadian, S., & Radisic, M. (2017). Nanotoxicity. In M. Razavi, & A. Thakor (Eds.). *Nanobiomaterials science, development and evaluation* (pp. 233-248). Elsevier. https://doi.org/10.1016/B978-0-08-100963-5.00012-4

Armarego, W. L. F. (2022). Nanomaterials. In *Purification of laboratory chemicals. Part 2 Inorganic chemicals, catalysts, biochemicals, physiologically active chemicals, nanomaterials* (9th ed., pp. 586-630). Butterworth-Heinemann.

AshaRani, P., Mun, G. L. K., Hande, M. P., & Valiyaveettil, S. (2009). Cytotoxicity and genotoxicity of silver nanoparticles in human cells. *ACS Nano, 3*(2), 279-290. https://doi.org/10.1021/nn800596w

Baig, N., Kammakakam, I., & Falath, W. (2021). Nanomaterials: A review of synthesis methods, properties, recent progress, and challenges. *Materials Advances, 2*(6), 1821-1871. https://doi.org/10.1039/D0MA00807A

Beaudrie, C. E. (2011). Using expert judgement for risk assessment. In A. Aschberger & C. E. Beaudrie (Eds.). *Assessing nanoparticles risks to human health* (pp. 109-138). Elsevier. https://doi.org/10.1016/B978-1-4377-7863-2.00005-4

Boehmke, B., & Greenwell, B. M. (2019). *Hands-on machine learning with R*. Chapman & Hall/CRC.

Concu, R., Kleandrova, V. V., Speck-Planche, A., & Cordeiro, M. N. (2017). Probing the toxicity of nanoparticles: A unified in silico machine learning model based on perturbation theory. *Nanotoxicology, 11*(7), 891-906. https://doi.org/10.1080/17435390.2017.1379567

Cunningham, P., & Delany, S. J. (2007). *k-Nearest Neighbour Classifiers.* Technical Report UCD-CSI-2007-3. https://researchrepository.ucd.ie/server/api/core/bitstreams/2de9257f-8a20-4b34-89f0-020a0183c7b5/content

Erkekoglu, P., & Kocer-Gumusel, B. (2018). Toxicity assessment of nanopharmaceuticals. In A. M. Grumezescu (Ed.), *Inorganic frameworks as smart nanomedicines* (pp. 565-603). Elsevier. https://doi.org/10.1016/C2016-0-04156-6

Frick, H., Chow, F., Kuhn, M., Mahoney, M., Silge, J., Wickham, H., & Posit PBC. (2023, August 8). *rsample: Part of the tidymodels framework*. https://rsample.tidymodels.org

Gajanan, K., & Tijare, S. (2018). Applications of nanomaterials. *Materials Today: Proceedings, 5*(1), 1093-1096. https://doi.org/10.1016/j.matpr.2017.11.187

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *International Journal of Intelligent Technologies and Applied Statistics, 11*(2), 105-110. https://doi.org/10.6148/IJITAS.201806_11(2).0003

Hansch, C., Maloney, P. P., Fujita, T., & Muir, R. M. (1962). Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature, 194*, 178-180. https://doi.org/10.1038/194178b0

Kou, L., Bhutia, Y. D., Yao, Q., He, Z., Sun, J., & Ganapathy, V. (2018). Transporter-guided delivery of nanoparticles to improve drug permeation across cellular barriers and drug exposure to selective cell types. *Frontiers in Pharmacology, 9*, Article 27. https://doi.org/10.3389/fphar.2018.00027

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5), 1-26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., Wickham, H., Hvitfeldt, E., Posit Software, & Posit PBC. (2024, February 18). *Preprocessing and feature engineering steps for modeling*. https://recipes.tidymodels.org

Li, J., Wang, C., Yue, L., Chen, F., Cao, X., & Wang, Z. (2022). Nano-QSAR modeling for predicting the cytotoxicity of metallic and metal oxide nanoparticles: A review. *Ecotoxicology and Environmental Safety, 243,* Article 113955. https://doi.org/10.1016/j.ecoenv.2022.113955

Misra, R., Acharya, S., & Sahoo, S. K. (2010). Cancer nanotechnology: Application of nanotechnology in cancer therapy. *Drug Discovery Today, 15*(19-20), 842-850. https://doi.org/10.1016/j.drudis.2010.08.006

OECD. (2014). *Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models* (OECD Series on Testing and Assessment No. 69). OECD Publishing. https://doi.org/10.1787/9789264085442-en

Pedersen, T. L. (2025). *patchwork: The composer of plots* (R package version 1.3.2.9000). https://patchwork.data-imaginist.com

Phanus-umporn, C., Shoombuatong, W., Prachayasittikul, V., Anuwongcharoen, N., & Nantasenamat, C. (2018). Privileged substructures for anti-sickling activity via cheminformatic analysis. *RSC Advances, 8*(11), 5920-5935. https://doi.org/10.1039/C7RA12079F

Puzyn, T., Leszczynska, D., & Leszczynski, J. (2009). Toward the development of "Nano-QSARs": Advances and challenges. *Small, 5*(22), 2494-2509. https://doi.org/10.1002/smll.200900179

Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X., Dasari, T. P., Toropov, A., & Leszczynska, D. (2011). Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nature Nanotechnology, 6*(3), 175-178. https://doi.org/10.1038/nnano.2011.10

Roco, M. C., Williams, R. S., & Alivisatos, P., National Science and Technology Council, & Committee on Technology, Interagency Working Group on Nanoscience, Engineering and Technology. (1999). *Nanotechnology research directions: IWGN workshop report*. IWGN.

Wen, Y., Lin, W., Mettenbrink, E. M., DeAngelis, P. L., & Wilhelm, S. (2021). Nanoparticle toxicology. *Annual Review of Pharmacology and Toxicology, 61,* 269-289. https://doi.org/10.1146/annurev-pharmtox-032320-110338

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., … Yutani, H. (2019). Welcome to the Tidyverse. *The Journal of Open Source Software, 4*(43), Article 1686. https://doi.org/10.21105/joss.01686

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x